FUNCTIONAL REQUIREMENTS FOR FIGHT DELAY PREDICTION

Data-Driven Probabilistic Flight Delay Predictions

In this section, we obtain probabilistic flflight delay predictions using two machine learning algorithms, Mixture Density Networks and Random Forests Regression.

Data Description:-

For this analysis, flight schedules available at Rotterdam The Hague Airport (RTM) between 1 January 2017 and 29 February 2020 are considered. In total, 17,365 departing and 17,336 arriving flights are considered. These flights arrive from and depart to 42 airports across Europe and North Africa. The shortest route included is to London City Airport(LCY), and the longest to Tenerife South Airport (TFS), with an average of 1300 km. shows a map indicating all airports to or from which flights depart or arrive. The delay distribution of these flight. The departing flights have an average absolute delay of 17.8 min with a standard deviation of 25.1 min, and the arriving flights have an average absolute delay of 15.4 min with a standard deviation of 26.4 min. Here, the delay is considered to be the positive or negative time difference from the schedule.

Weather Dataset:-

We also consider the weather conditions, such as the temperature, pressure and wind speed, measured at the origin/destination airport of all flights arriving/departing at RTM in the period 2017–2020. Measurements are available every 30 min.

Feature Selection :-

Feature selection is performed using the Pearson Correlation Coeffificient .The correlation between any two features and the correlation between the features and the target (the flight delay) are calculated for a given training set. The features are selected as follows: for any two features are correlated by more than the threshold value of 0.7, the feature that has the smallest correlation with the target variable is removed. features that have been selected for flight delay prediction a

description is provided for each of the selected features. The features Airport, Airline, Season, Time of day, Day of week, Day of month, Day of year, Airport latitude and longitude, Distance, Month, Year and Scheduled flflights 2h and day are obtained or calculated from the flight schedule dataset. The feature Seats is derived from the aircraft type assigned to perform a flight. The features Temperature, Dewpoint, Visibility, Pressure, and Wind speed are obtained from the weather dataset

Prediction Features:-

Departure delay Airport a , Airline a , Season a , Time of day b , Day of week b , Day of month b , Day of year b , Airport latitude c , Airport longitude c , Day of month c , Seats c , Year c , Scheduled flights 2 h c , Scheduled flights day c , Dewpoint c , Visibility c , Pressure c , Wind speed c .Arrival delay ,Airport a , Airline a , Aircraft type a , Season a , Time of day b , Day of week b , Day of month b , Month b, Airport longitude c , Day of month c , Distance c , Seats c , Year c , Scheduled flights 2h c , Scheduled flights day c , Temperature c , Visibility c , Pressure c , Wind speed c. This feature is target encoded; b This feature is trigonometrically encoded; c This feature is numerically encoded.

Feature Description:-

Airport - the airport of destination (departures) or origin (arrivals)

Airline - the airline operating the flight

Aircraft - type the aircraft type used for the flight

Season - the flight season (summer or winter schedule)

Time of day - scheduled time of day of the flight

Day of week - scheduled day of the week of the flight

Day of month - scheduled day of the month of the flight

Day of year - scheduled day of the year of the flight

Month - scheduled month number of the flight

Airport - latitude and longitude the latitude and longitude of the destination/origin airport

Distance - the distance between the origin and destination

Seats - the seat capacity of the used aircraft

Year - the year in which the flight was operated

Temperature - the air temperature at the destination/origin airport

Dewpoint - the dewpoint temperature at the destination/origin airport

Visibility - the prevailing visibility at the destination/origin airport

Pressure - pressure altimeter at the destination/origin airport

Wind speed - wind speed at the destination/origin airport

Scheduled flights day - the number of flights scheduled to depart/arrive during the day of flight

Scheduled flights 2h - the number of flights scheduled to depart/arrive during the period between one hour before and one hour after the scheduled time of the flight.

The features are either categorical, time-related, or numerical. The categorical features are target encoded based on a binary delay threshold of 15 min. The encoded value of the sample feature is the delay rate of the category to which the sample belongs. For example: if 8 out of 20 samples flying on Tuesdays are more than 15 min delayed, all Tuesday flights are encoded with value 0.4 for the feature Day of the week. The time features are encoded using trigonometric functions to preserve the periodicity. Two features (sine and cosine) are extracted from every time feature. For example, the features Month sine and cosine are calculated using $\sin(2\pi m 12)$ and $(\cos 2\pi m\ 12)$ for a given month m.

The remaining features are numerically encoded, i.e., the encoded value is the same as the original feature value. Note that the time features are both trigonometrically and numerically encoded. For example, the data field Day of the week yields the features Day of the week sine, Day of the week cosine, and Day of the week. The encoding method of every selected feature is denoted . After encoding, all feature values are scaled to the interval [0, 1] to eliminate undesired feature domination in neural network classifiifiers most features are selected for at least one of the departure/arrival pair, and that the trigonometrically encoded time features are selected more often

than the non-encoded time features.

Machine-Learning Algorithms to Estimate the Probability Distribution of Flight Delays:-

Two algorithms are proposed to estimate the distribution of flflight delays: Mixture Density Networks (MDN) and Random Forests regression (RFR). These algorithms belong to different classes of machine learning algorithms, neural networks, and decision trees, respectively. Mixture Density Networks (MDNs). A Mixture Density Network is a combination of a neural network and Gaussian mixture model. Given feature values $\mathbf{x}i$ of flflight i, an MDN outputs the parameters for each Gaussian in the mixture: the weight α, the mean μ, and the standard deviation σ.

With these parameters, the probability density function $p(yi|\mathbf{x}i)$ of the target variable yi, the flight delay, is determined. In general, the MDN is particularly suitable to estimate multimodal probability distributions. It is therefore able to predict a distribution with peaks at, for example, two separate likely delay values.

The flight delay probability distribution is constructed as the weighted sum of Gaussian distributions as follows:

$$p(yi|\mathbf{x}i) = m\sum j=1\ \alpha j(\mathbf{x}i)\varphi j(yi|\mathbf{x}i), \quad (1)\ \varphi j(yi|\mathbf{x}i) = 1\ \mathrm{q}\ 2\pi\sigma j(\mathbf{x}i)2\ \exp -(yi - \mu j(i))2\sigma j(\mathbf{x})2\ !\ (2)$$

where $p(yi|\mathbf{x}i)$ is the probability distribution of delay value yi given feature values $\mathbf{x}i$ from flight sample i, while $\alpha j(\mathbf{x}i)$, $\mu j(\mathbf{x}i)$ and $\sigma j(\mathbf{x}i)$ are the weight, mean, and standard deviation of the jth Gaussian component, 1 components considered for the mixture. $\leq j \leq m$ with m the total number of Gaussian while the parameters αj, μj , and σj are the output of the MDN. Thus, there are $3m$ outputs of the MDN. The weights use a softmax activation function, and the standard deviations use an exponential activation function, while the means are unrestricted. The neural network is trained using backpropagation, i.e, the network parameters, the weights and biases of each node are updated using an error function E, which is the negative logarithm of the likelihood that the model derived from the output of the current network gives rise to the training data . This likelihood is the product of the likelihood of every data point, given the current network parameters.

$$E = Nf\ \sum i=1-\ \ln\ m\sum=1\ \alpha j(xi)\varphi j(yi|xi)! = Nf\ i=1\ -\ \ln\ p(yi|xi),\ (3)$$

where Nf is the total

number of samples in the training set. For every data point fed to the neural network, the derivatives of the error with respect to all network parameters are used to update the weights and biases of the network. Following training, the MDN is applied to a test set and multimodal probability distributions for the delay of each flflight in the test set are estimated. The MDN method illustrated. Schematic representation of a Mixture Density Network: parameters for a multimodalGaussian distribution are obtained using a Neural Network. Random Forests Regression and Kernel Density Estimation Random Forests regression (RFR) is a class of decision tree-based machine learning algorithms . The regular RFR algorithm is an ensemble method that combines the results of a number of decision trees. When building each tree, a random subset of the feature values of each training data point is used to make branches. The algorithm outputs a point estimate for the target variable (flflight delay) of every test sample by averaging the output values of all considered decision trees. However, for our analysis, we are interested in estimating the probability distribution for the delay of the given flflight, rather than a point estimate. In order to obtain the flight delay distribution of a flflight in the test phase, the output values of the decision trees are not averaged, but collected, and a kernel density estimation (KDE) is performed . A KDE results in a normalized probability density function. Two settings of the KDE are the kernel type and the bandwidth. In our analysis, a bandwidth of 1.5 is used to render the estimated distribution smooth. Gaussian kernels have been selected for their generality. Random Forests regression is a well -established technique that has been applied in many research areas. However, there are very few examples of studies utilizing the algorithm to obtain probability distributions. För steretal use quantile values, obtained from Quantile Random Forests, to construct a right-continuous cumulative distribution function of aircraft's time-to-flfly from the turn onto the fifinal approach course to the runway threshold. Schlosseretal and Rahmanetal use Random Forests algorithms to obtain probability distributions for precipitation forecasts and drug sensitivity, respectively.

Both studies make use of feature probability distributions estimated via maximum likelihood to make splitting decisions when constructing the decision trees. In contrast,

in this study, the feature values and splitting decisions are kept deterministic throughout the Random Forests algorithm. In this way, the probability density function is estimated from deterministic feature values without the need for stochastic variables. Furthermore, the working of the original Random Forests regression algorithm need not be changed.

Hyperparameter Tuning:-

The hyperparameters of the MDN and the RFR prediction algorithms have been optimized using a grid search. The hyperparameters leading to the lowest mean CRPS scores have been selected shows the selected hyperparameters and their search range. For MDN, a network with three hidden layers of 50 nodes is selected. The output layer of the network consists of 24 nodes, with which an 8-modal Gaussian distribution function is constructed. For RFR, 200 decision trees with a maximum depth of 10 layers are constructed. For every branch split, three out of four features are considered of at least seven training samples.

MIXTURE DENSITY NETWORK :-

Number of modes m = 8 = [3, 5, 8, 10, 15]

Number of hidden layers = 3 = [1, 2, 3]

Number of nodes per hidden layer = 50 = [25, 50, 75, 100]

Number of epochs = 1000 = [500, 750, 1000, 1250, 1500]

RANDOM FOREST REGRESSION:-

Number of estimators = 200 = [100, 150, 200, 300]

Split criterion = Mean-squared error = [MSE, MAE]

Maximum tree depth = 20 = [4, 6, 8, 10, 12, 15, 20, 30]

Minimum samples per leaf node = 7 = [0, 3, 5, 7, 9]

Fraction of features considered for split = 0.75 = [0.25, 0.50, 0.75, 1.00]

KDE Bandwidth h = 1.5 = [0.5, 1, 1.5, 2]